A Novel Approach for improving Prediction in Data Mining

SHORT SYNOPSIS

Submitted in partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

Manav Rachna University

by

Ms. Pinki Sagar
Enrollment number: 2K16-PFE-018

Under the supervision of

Dr. Prinima
Associate Professor
Department of Computer Science & Technology
Faculty of Engineering

Co-Guidance
of
Dr. Indu Kashyap
Professor
Department of Computer Science & Engineering
Faculty of Engineeringand Technology

MANAV RACHNA
vidyanatariksha
MANAV RACHNA UNIVERSITY
(FORMERLY MANAV RACHNA COLLEGE OF ENGINEERING
NAAC ACCREDITED 'A' GRADE INSTITUTION)
Declared as State Private University under section 2f of the UGC act, 1956

Year 2019

MANAV RACHNA UNIVERSITY

Established, vide Haryana Act no.26 of 2014

(Formerly Manav Rachna College of Engineering,

NAAC Accredited 'A' Grade Institute)

DECLARATION BY THE SCHOLAR

I declare that the Synopsis entitled "A Novel Approach for improving Prediction in Data Mining." is prepared under the supervision of Dr. PRINIMA GUPTA and Dr. INDU KASHYAP and submitted for the approval of Departmental Research Committee (DRC).

16/1/19

Dr. Prinima Gupta

Signature of Supervisor

Dr. Indu Kashyap

Signature of Co-Supervisor

16/1/19

Pinki sagar

Signature of scholar

Proposed Title of Thesis: A Novel Approach for improving Prediction in Data Mining.

Abstract: Data mining is the set of algorithms deliberate to find the hidden knowledge from the data sets, some of the popular techniques of data mining are prediction, sequential patterns, association, classification, clustering, and decision tree. Data mining techniques are categorized into two categories: Predictive and Descriptive techniques. Regression and classification are the predictive mining techniques. Association and clustering are the descriptive mining techniques. Classification and regression are used for forecasting. Regression algorithms are based on various regression model i.e. linear regressions, non-linear regression, multiple regressions, logistic regression, and probabilistic regression. Regression techniques have an important role in the prediction of continuous data, which can be frequent item stream data or time series data. It can be one dimensional or multidimensional stream data. Regression based techniques for prediction of stream data is frequent item prediction method (FIPM), frequent temporal pattern data stream (FTPDS), Sequence frequent stream data for plane regression stream (SFAPR).SFAPR overcome the problem of FIPM and FTPDS. Forecasting algorithms have important role in the area of traffic, weather forecasting, business analysis, share market, stock exchange, mobility, medical etc. In some area like drilling, stock exchange, weather forecasting prediction can help in increasing the revenues and performances. It can also be done by combining classification and regression techniques. In the proposed work forecasting of time series data sets with improved parameters is done. For preprocessing the data set, sliding window or classification algorithms are used. Then coefficients values for the regression model are identified to fit the regression model. It can be used to inspect the characteristics of the data.

Results: accurate prediction, reduction of standard errors, data cleaning. To improve the linear models quadratic, root or polynomial functions can be used. Evaluations of regression models are done to check the standard errors or residual plots so that it can help to avoid inadequate models and help to fitt the model for better results.

I Introduction

Databases are growing day by day and it is very difficult to extract useful information by using SQL and other tools from these large databases. For retrieving data from the (usually unstructured or poorly structured) large data sets is a process of mining or extraction of data. Extracted data is used for data processing, data analysis or taking decisions.

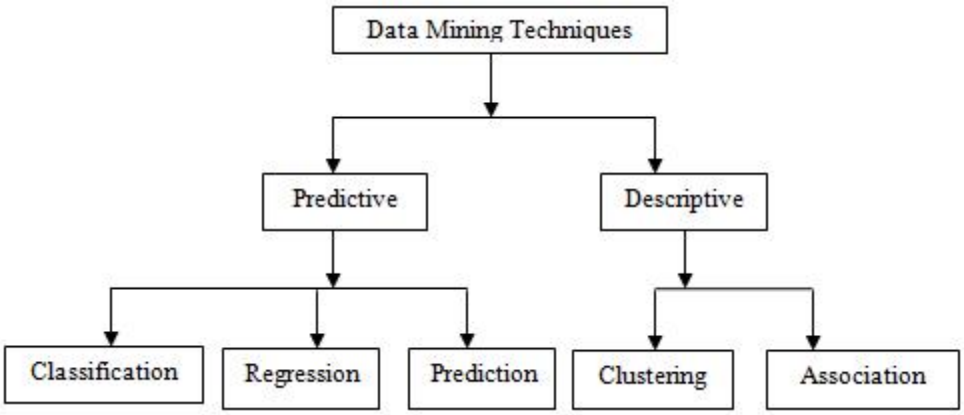


Figure1: Data Mining Techniques.

All data mining methodologies are divided in to two major categories Predictive and Descriptive mining. Descriptive Mining: method of data aggregation and data mining to provide insight into the past and its answer: "What has happened?" Predictive Mining: statistical or arithmetical methods and prediction techniques are used to understand the future and it answer: "What could happen?"In data mining many techniques are used and each has its own application in various areas. Regression and Classification are the predictive methods which are used for analysis, forecasting and extraction of data sets of data ware houses.

Two data mining techniques used for forecasting are as follows −

1. Classification
2. Regression

Classification: The methods of classification in data mining techniques can predict data like 0 and 1 which is known as categorical data; and prediction models predict continuous valued functions. Considering an example, of loan applications in Bank, by applying classification algorithms we can categorize the bank loan application which is safe or not. The big task is how to preprocess data so that data sets can be prepared for Classification and Prediction and we can apply the algorithm of data mining techniques.

Regression: In Data mining forecasting based on historical data is called prediction. It could only be done by considering the facts or evidences. In basic prediction technique (linear regression), can only predict values of one variable which is known as dependent variable from the value on a second variable, which is termed as independent variable. The model for linear regression is Y=a+b*X, where Y is referred as criterion variable as an independent variable and X is referred as base (independent variable) for predictions which is termed as the predictor variable. If one predictor variable is available, then the way of prediction is known as simple linear method of regression techniques. In the basic linear regression method, the predicted values of Y(dependent variables) are plotted which are to be find using linear regression model and value of X and make a straight line, finding the best fitted straight line through the predicted values of Y is called as Linear regression. This best-fitted straight line is for showing regression analysis known as regression line and is also plots the predicted values on Y for each possible value of X.

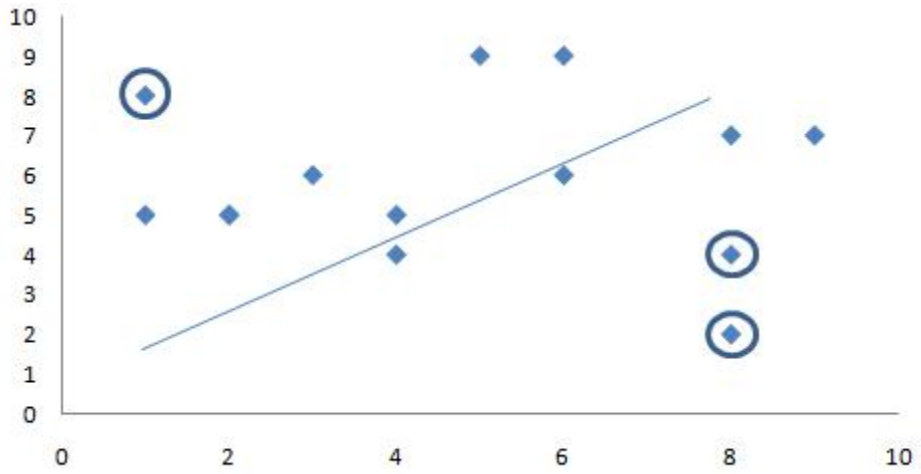


Figure2: Outliers in regression model.

As we can see in figure 2, some plotted values are very near to the regression line; it means that there is less errors in prediction and some plotted values are very far away from the regression line; it mean there are more errors in prediction. So we can do analysis that our prediction is with high errors or low errors by using linear regression technique.

Comparative Analysis of Classification and Regression:

Parameters	Classification	Regression
Performance	Satisfactory	Satisfactory
Technique Complexity	Satisfactory	Very High

Requirement of Dependent Variable	No	Yes
Algorithms	ID3 algorithm,C4.5 algorithm,SLIQ algorithm	FIPM, FTPDS, Linear, Nonlinear regression
Type of data	Two dimensional data	1-D and 2-D data
Applications	Expert Systems and statistics Neuron Biology	Deviation Detection
Analysis	Discriminate	Calculation of errors during the prediction
Supervised /Unsupervised learning	Supervised	Supervised

II Problem Statement

In the earlier research forecasting through regression model for time series data sets can help in understanding of prediction and the behavior of real time systems from experimental or observational data. Some prediction models are not certain due to the low reliability of sources or lots of problem used for measurements and observation. In the proposed method the prediction model will improve the accuracy of prediction rate and reduce the standard error of the time series dataset by using one of the best regression models.

III Methods for Regression Analysis of Stream Data

Stream data is continuous data; it is collected with fixed time interval in time. It can be one dimensional and two dimensional stream data. Below are the techniques for predicting stream data:

1. FIPM: - (Prediction Method for Frequent item data sets).
2. FTP_DS:- (Prediction Method for Frequent Temporal Pattern Data Stream)
3. SFA-PR:-Sequence Forecast Algorithm for plane Regression

Frequent Item Prediction Method (FIPM): The method used for analysis of stream data using regression (one dimensional). Prediction of stream data is to be done by linear regression model. First the data would be preprocessed and convert it in the appearance of sample value for using in further techniques of prediction using regression. When sample results are produced, a regression analysis based model is generated that is least squares method, it is a method to finding of regression model. The objective of analysis method is to find a predictor through forecasting

method from samples which are experiential values. So, it is required to convert the (already accured) frequent stream data into a pair, such as (Xi,Yi).

The key characteristics of Frequent Item Prediction Method are the following:

1. Formulate the data according to the accruing of time, when similar types of data are occurred.
2. Find out the difference of time intervals when particular variables are occurred.
3. Form the pairs using intervals in the form of (X, Y).Make the pairs in the form of two type variable these are: dependent variables and independent variables.
4. Fit the data in to linear regression model (preprocessed).

Frequent Temporal Pattern Data Stream (FTPDS): This algorithm was introduced by Wei-Guang Teng in 2003. There was two important attributes for Frequent Temporal Pattern Data Stream (FTPDS) algorithm, first is the frequent occurrences of data sets sequence and its support values and second is the pattern representation.

The support is defined as frequent accruing sequence (a) at a particular time (t) which is defined by the proportion of the number of events have sequence (x) in each sliding window which is in currently used, to total number of events.

The key characteristics of Frequent Temporal Pattern Data Stream are the following:

1. This algorithm checks the flow of transaction (online) and generates frequent patterns for real time data.
2. Second significant characteristics of Frequent Temporal Pattern Data Stream (FTP-DS) compact pattern representation based on regression techniques. Particularly, to meet up the space restriction, partition for pattern representation a compressed Accumulated Frequency and time form to aggregately include all the data and information which is necessary for further regression study for prediction.
3. In comparison of other linear regression model .FTP-DS perform the identification of errors during prediction is very effectively.
4. FTP_TS algorithm is used for prediction for multi dimensional data(stream data

   Two dimensional).

Sequence Forecast Algorithm for plane Regression (SFA-PR): Sequence Forecast Algorithm for plane Regression calculates support for particular sequence and then applies the plane

equation to predict the sequences or assurances of data sets in the future. In comparison with the other mining algorithms for predicting of sequences, Sequence Forecast Algorithm for plane Regression SFA-PR can include more area and never forget key exceptions.

IV. **Review of Literature**

In data mining significant patterns and hidden relation between attributes from large amounts of data are automatically determined. Prediction techniques of Data mining identify correlations between variables. Linear regression is used in any system for forecasting one dimensional and two dimensional data.

Xiao ZhouKiang-Yu Liu, Stephen T.C. 2004 [1] Wong, explains that logistic regression is a specific model for designing of prediction model for categorical data or binary data like 0 or 1. The result of logistic regression has been improved for classification of disease and data sets for disease in the form of microarray. A characteristic selection step should be added to penalize logistic regression due to a large number of features and a less number of subjects.

FENG ZHAO, QING-HUA LI 2005 [2] proposed a method which is based on plane regression. SFA-PR predicts the sequences of two dimensional frequent stream data. After collecting and integrating of real-time frequent stream data it is preprocessed by using of sliding window algorithm. Proposed algorithm finds frequency for occurrence of sequence and support for described sequence then applies the plane regression to predict the sequence trends in the future. SFA-PR mining algorithm gives better result in comparing with other prediction algorithms which are used in forecasting of sequence data sets. SFA-PR can cover many fields and never neglect th exceptions.

Duck JinChai, EunHee Kim and Long Jin 2007 [3] introduced a method (FIPM) for the prediction of one dimensional frequent stream data sets. Frequent Stream data has characteristics of continuous and infinite. Frequent Item Prediction Method can forecast frequent stream data items by using basic regression (linear) method for one Dimensional stream data. Historical data (time dependent) helps in predicting of future through the regression analysis. In the Frequent Item Prediction Method, first data preprocessing is applied to form data attributes in the form of coefficients that are used in regression analysis and then apply simple linear regression model.

Andreas Hecker, Thomas K¨urner2007[4], An Automatic method is introduced for the location area code planning prediction values of the probable signaling traffic are required for

mobile networks. Traffic and mobility needs provide Data. The method is based on knowledge of data sets. These data sets are related to geographical information such as population distribution, and land use which is considered as value of data input. The output data is performance values measured from the(OMC) Operations and Maintenance Center. On the basis of classification and regression trees basically author estimate the consummation analysis of fast modeling methods. The analysis of consummation is evaluated out by comparative analysis the values of Mobile Terminated Call with OMC (operations and maintenance center) measurement values from a real network.

Qingshan Ni1, Zhengzhi Wang, Qingjuan Han 2009 [5], Introduced novel approach based on the logistic regression, to predict protein function from protein-protein interaction data. Experimental Result shows that the proposed method can obtain advantageous results for protein function prediction.

Alberto Landi, Paolo Piaggi2010, [6] apply the regression and classification techniques for the data analysis of continuous and categorical data. These techniques are most commonly used in analysis. So that, a number of optimization process begin from various required initial parameters which were selected randomly, and the optimum execution is selected as the result. In addition, trial-and-error approach helps in finding of the number of hidden neurons in the artificial neural network: trial-and-error approach adopt the techniques such as generalized cross validation and information criteria can provide the upcoming guidance on the selection of the most suitable artificial neural network architecture.

SonaliTiwari 2010 [7] introduced a method for improving prediction of two dimensional frequent stream data FTP_DS algorithm with non linear regression, so the more exceptions can cover by sequence trends. In this paper, comparison of two algorithms: FTP-DS based on linear regression and FTP-DS based on nonlinear regression algorithm is performed. An obtained result through the proposed method improves the error rate during the prediction.

Mythili T., DevMukherji, Nikita Padalia, and Abhiram Naidu [12] has introduced method to point out the heart disease factors as they will perfectly forecast the risk factors for heart disease using methods of data mining . They proposed a model that is rule based classification is used to analyze the accuracy of the results in decision trees, rules based, (SVM) support vector machine, and logistic regression. Data mining is the process of taking out the needful information from the

huge data sets which results in the prediction of the data by using different methods of data mining such as regression, classification, clustering, association, etc. They gave a proposed structure using combination of logistic regression, (SVM)support vector machines, and method decision trees helps to get a correct forecasting of heart disease.

Nirmala, M. and V. Palanisamy 2013 [13] emphasizes on finding of weak students or slow learners. With the help of classification algorithms result is predicted. Real data set is collected from a high school and filtration of required potential variables is performed using WEKA an Open Source Tool.

Bolin Chen, Min Li y, Jianxin Wang y and Fang-Xiang Wu, 2014 [16], in their words it is clear that recognition of ailment genes is the initial step to understand the genetic disease mechanism. Even though lots of computational techniques are introduced to recognize ailment genes, they have low performance in terms of AUC scores or are taking a lot of high time. To control these problems, regression based algorithm are used (such as logistic regression) in their study for identifying of ailment genes. The problem of ailment gene detection is framed in two classes of classification techniques, where one class is ailment genes, whereas another class has non-disease genes. Logistic regression is based on binary values such as 0 or 1 and it is used to forecast the chances of a gene related with ailment by considering the prior information as the dependent variables in various category and label which are related to feature vectors termed as estimator variables. Numerical practical's illustrated that the proposed regression ((based on binary data) logistic) based algorithm has a very fine performance and noticeably reduce the execution time.

M.A.NisharaBanu, 2014 [17] introduced a method to predict the heart disease through classification algorithms. In medical area concealed data sets patterns could be used for the study of disease related to heart. Proposed approach is successful approach for the most recent and imprecise patterns in the real data sets. The informed data provided by the hospitals can give the result for better prediction in the field of heart disease.

Aida Mustapha ,Farhana M. Fadzil, 2015 [18]introduced a regression strategy to predict the annual profit of vendors based on their previous data so the evaluation exercise could be prepared .In This paper author computes vendors' performance by using three different regression models: Linear Regression, Sequential Minimal Optimization algorithm and M5rules. The results

shows that the ability of data mining tasks in helping Entrepreneur Development Unit (EDU)to predict vendors' performance and to identify groups of on performance and under-performance.. (EDU) is responsible to manage a big group of vendors that hold contract with the company.

GustiAyuPutriSaptawati ,GustiNgurah Mega Nata 2015 [19],states that the major activity in mineral industry which is most significant and costly is drilling. All though, finding of targets for drilling, geologist are using qualitative study, which results in lot of failures in drilling. The authors work mainly emphasis on the analysis of methods, used for mining, and used to retrieve the facts from this data. In the final outcome, the categorization of informed data and mining of common item sets are able to maintain the forecasting of drilling targets. The objective of work is to reduce the threat of failure of drilling and hold the industry's decision and deciding a new target in drilling.

Sheela Gole 2015 [21] says terrific quantity of data is being explore through internet and from various sources for example social media feedback, applications of internet etc, and data collected from these types of resources. Data could be managed by traditional tools and techniques.

Swati Gupta 2015 [22, 21] discussed about the framing of linear regression technique, with the designing of linear regression method, training data sets or test data sets are used to fit a equation of linear regression. It is the very simplest method for regression analysis, and find out the suitable values for coefficients a and b. The value of dependent variable y based upon a assigning value of independent variable X. Basic models of regression are used to estimate the association between two variables or factors. Linear regression model is represented by the following equation:

$$y = \alpha + \beta X$$

In mathematics there are two variables which are used in a simplest linear regression analysis or study, X and Y are the variable used in linear regression. The regression model explains how y is dependent on x or how they are related is recognized as the linear **regression model. I**n the equation which is defined in above, α is the y intercept of the regression line and β is the linear slope.

Ilayaraja M, Meyyappan 2015, [23] define the techniques of Data mining are applied in many area of medicine for different objectives. They partitioned a process to estimate the risk factors of the patients who's having symptoms of heart disease through the collected frequent data sets.

Data sets for the heart patients are collected from the medical institutes or hospitals. Frequent data item sets are produced and depended on the selected symptoms and minimum support value. The frequent or common data sets which are extracted could help the doctors to take decisions in diagnostics and to predict the level of risk for at an early stage so that immediately treatment can be provided. The projected approach could be applied to data set of medical field which helps in predicting of the factors that affect risk with level of risk and the patients based on selective factors item sets.

Jae Kwon Bae and Jinhwa Kim2015 [24], introduce a method called Ubiquitous Data Mining that actually consolidates different uniform models in a much smart computing background. Based on this author introduce many personal credit rating prediction models and marks their performance to compare with other models such as Logistic Regression (LR), Multilayer Perceptron (MLP), Bayesian style frequency matrix (BFM), classification tree methods (C5.0) and neural network rule extraction (NR) algorithms.

Farhan Khan, Dariush Kari, IlyasAlper Karatepe, and Suleyman S. Kozat2016 [25], author define non-linear regression by assuming that the data depend on a various fold. They divide the data space into multiple areas to construct a partitioned linear regression analysis as an estimation of the non-linearity among the experiential and the expected data, in place of setting up the range and limitations of the particular category, Algorithm is exposed to immediate adapt to the differences in the data and it is very successful for high dimensional data with small data sets.

YoungjoLee2017 [26] Gaussian process regression model mostly used to fit data set when unknown regression function is used. In this paper author introduce a novel approach, Extended - Process Regression model, a nonlinear regression model which allows a robust Best Linear Unbiased Predictor by applying concise construction. it over take many compelling properties from the Gaussian process regression GPR model.

V Literature Conclusion

In literature survey it has been studied various prediction techniques by using regression and classification for different types of data. Data could be stream data, time series data frequent data .Three algorithms used for forecasting are, FIPM,FTPDS,SFAPR In earlier FIPM (based on

linear regression) ,Linear Regression is to be applied for the frequent one dimensional stream data and it produce the Prediction graph with high errors and low accuracy. Stream data mostly contains incomplete data items, lots of noisy and overflowing data items. (Frequent Item Prediction Method)FIPM and (Frequent Temporal Pattern Data Stream) FTP-DS are the two algorithms which are used for mining of multi dimensional or one dimensional stream data.In comparative study of these two algorithms it has to be identified that these two algorithms use the linear regression method and gave value of estimator with high error. In non linear regression based FIPM and FTPDS we get the improved results but in FIPM based on non linear regression but did not generate the much good result in reduction of errors. Frequent Temporal Pattern Data Stream (FTPDS) using non linear regression gives forecast with improved errors rate or we can say that low error in comparison of linear regression based FIPM and FTPDS. The error detection rate find out by algorithm SFA-PR is much higher than method Frequent Temporal Pattern Data Stream. The SFA-PR helps to clarify or explain prediction more intruding than FPT-DS algorithm so that it covers each and every data set. Logistic regression is based on binary values such as 0 or 1 and it is used to forecast the chances of a gene related with ailment by considering the prior information as the dependent variables in various category and label which are related to feature vectors termed as estimator variables. Regression strategy can be used in predicting the annual profit of vendors based on their previous data so the evaluation exercise could be prepared. In some forecasting methods regression and classification techniques are combined. The forecasting approach could be applied to dataset of medical field which helps in predicting of the factors that affect risk with level of risk and the patients based on selective factors item sets.

VI Objective of Proposed work

Regression is used for prediction of different types of data sets. In a multivariate data set if we apply the prediction for a multivariate data set, all the independent variables are not necessary to include in the prediction model. To increase the prediction accuracy rate of multivariate data sets irrelevant variables are not considered as independent variables. The following research objectives have been achieved for the completion of research work:

RO1: Study and analysis of existing regression techniques.

RO2: Selection of Significance of independent variables for the prediction.

RO3: Design a novel prediction algorithm based on regression.

RO4: Implement a proposed prediction algorithm and compare it with existing algorithms.

VII Methodology

Methodology of Regression is a quantitative research method which involves modeling and analyzing many variables, where the relationship includes a dependent variable and independent

variables (one or more).

STEP1: Collection of data is done from online repository (UCI Repository).

- Characteristics of data set
- Data type: time series data set
- Attribute characteristics: Real
- Associated task: Regression
- Source: UCI repository

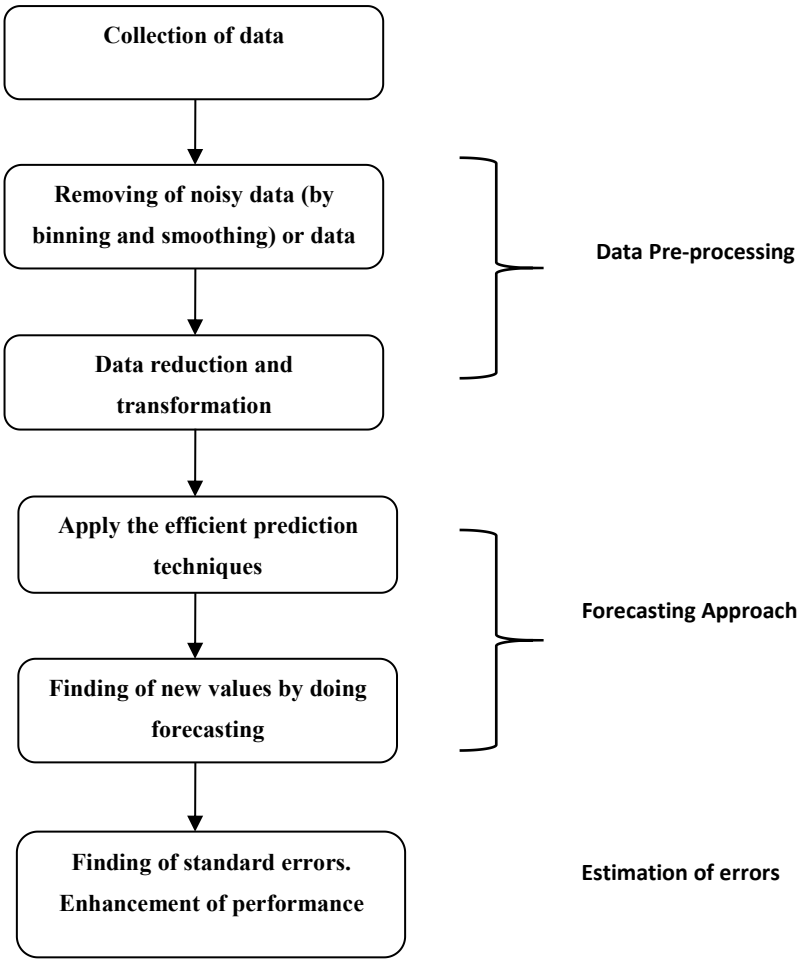
STEP2: An outlier or an error is an observation that lies an irregular distance from other values in a sample data (randomly selected) from a population data It means that, the analyst decide what should be considered and should not be considered. Before unusual observations could be selected out, it is mandatory to find the feature of normal observations.

STEP3: Data preprocessing is a step of data mining technique that involves conversion of raw data into a model so that regression could be applied or an understandable form. Real-time or dynamic data is often not complete, not consistent, deficient in certain behaviors or sequence trends, and it may have contain many errors. Huge data sets are usually incomplete due to several reasons, absence of attribute values, incorrect data, non relevant data and corrupted data. Containing errors or outliers.

- **Data cleaning**: It is the method of removing of corrupt or inaccurate data from a data sets, and refers to identifying imperfect, wrong, erroneous or irrelevant parts of the data sets and then updating, or removing the dirty data.

  - It is the procedure of fill up the missing values or ignores the row, usually done when labeling of data is missing.
  - Fill the mean value in the place of missing value for all samples belongs to the same type of class.
  - Learning algorithms are used to predict the missing values. Consider the missing value of attribute as a dependent variable and run a learning algorithm to predict the data (missing values).

- Binning is the method which is used to Sort the attribute values and partitions them into different bins. Then smooth by bin means, bin median, or bin boundaries.


Collection of data
Removing of noisy data (by binning and smoothing) or data
Data reduction and transformation
Data Pre-processing
Apply the efficient prediction techniques
Finding of new values by doing forecasting
Forecasting Approach
Finding of standard errors. Enhancement of performance
Estimation of errors

- **Data transformation**: It is the practice of normalizing the data. It is the procedure of Scaling attribute values to fall within a particular range.

- **Data reduction**: It is the method of dropping the count of attributes by Data cube aggregation operations, slice or dice, roll-up, operations., filtering and wrapper method are attribute selection method for removing of non relevant attributes. Principle component analysis is a method for searching the lower dimensional space that can represent the data in best way.

STEP 4: Calculation of coefficients (depend on regression model).

- Linear Regression, co efficient are a0 and b1

$$Y = a_0 + b_1X_1$$

- Nonlinear regression uses following equation:-
  Y = f(X,β) + ε
  Where:

  X = a vector of *p* predictors,

  β = a vector of *k* parameters,

  ε = an error term.
- Multi Linear regression

  $$Y = a_0 + b_1X_1 + b_2X_2 + b_3X_3+\varepsilon$$

STEP 5: Apply the Regression model depend on type of data for example if they are using two dimensional data in number format they will use linear and nonlinear regression model. If processed data is in binary forms then they will apply logistic regression. And so on.

STEP 6: After applying regression model they can check the accuracy of prediction which is based on independent and dependent variable. Percentage of accuracy is found out using standard error. The standard error of the estimate is a measure of the accuracy of predictions.

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

Error= Actual value-Predicted value

STEP 7: During Prediction of data, standard error can be find and data analysis which is identified by difference of actual data and predicted data. Data Analysis is a method that help describe facts, test hypotheses and detect patterns.

- **Correlation Analysis**: Some Databases may have data sets that contain some irrelevant attributes. Finding two attributes of the data base are interrelated or not can be only analyzed using correlation analysis. In correlation analysis, we measure the relationship between more than one continuous variable (for example: between an independent and a dependent variable or between two independent variables).
- **Regression Analysis**: In regression analysis More than one variable are included, one is independent variable and second is dependent variable. In multi linear regression multiple

independent (more than one variable) are used. Regression analysis is the procedure of checking error rate during the prediction or it is a way of predictive modeling. It is used in the analysis of relationship between a target and predicator. It is also used in forecasting or predicting of the datasets based on previous data sets or historical data. Regression techniques are used in finding out the relationship between the dependent and independent variables in data sets. For example, relationship between road accidents and rough driving by a driver is best studied through regression(prediction model). Regression is a statistical process for estimating the relationships or correlation among different attributes of data sets. Analysis could be done through using many regression techniques for model and analyze different attributes, whether the main focus is on the correlation between dependent variable and more than one independent variables

References

[1] XiaoboZhou,Kuang-Yu Liua, b,Stephen T.C. Wonga, b259 Biomedical Machine Learning Cancer classification and prediction using logistic regression with Bayesian gene selection, Volume 37, Issue 4, August 2004.

[2]FENG ZHAO,QING HUA LI, A plane regression-based sequence forecast algorithm for stream data, **Published in:** Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on18-21 Aug. 2005 IEEE.

[3] Duck Jin Chai EunHee Kim Prediction of Frequent Items to One Dimensional Stream Data Fifth International Conference on Computational Science and Applications 0-7695-2945-3/07 © 2007 IEEE.

[4] Andreas Hecker, Thomas K¨urnerApplication of Classification and Regression Trees for Paging Traffic Prediction in LAC Planning 1550-2252/$25.00 ©2007 IEEE.

[5] Qingshan Ni1, Zhengzhi Wang1, Qingjuan Han2, Gangguo Li1, Xiaomin Wang1, Guangyun Wang1 Using logistic regression method to predict protein function from protein-protein interaction data 978-1-4244-2902-8/09/$25.00 ©2009 IEEE.

[6] Alberto Landi, Paolo Piaggi Artificial Neural Networks for Nonlinear Regression and Classification 978-1-4244-8136-1/10/$26.00_c 2010 IEEE.

[7] SonaliTiwari" Sequence Forecast Algorithm Based on Nonlinear Regression Technique for Stream Data" Vol. 1, No. 2, July-December 2010.

[8] KrisztianBuza, AlexandrosNanopoulos, Lars Schmidt-Time-Series Classification based on Individualised Error Prediction Thieme 2010 13th IEEE International Conference on Coputational Science and Engineering 978-0-7695-4323-9/10 $26.00 © 2010 IEEE.

[9] Umesh Kumar Pandey, Saurabh Pal Data Mining : A prediction of performer or underperformer using classification (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), 2011.

[10] NidhiBhatlaKiranJyoti An Analysis of Heart Disease Prediction using Different DataMining Techniques International Journal of Engineering Research & Technology (IJERT) vol. 1 Issue 8, October – 2012.

[11] Rita P. RibeiroTowards Utility Maximization in Regression 2012 IEEE 12th InternationalConference on Data Mining Workshops 978-0-7695-4925-5/12 $26.00 © 2012 IEEE

[12] Mythili T., DevMukherji, Nikita Padalia, and Abhiram Naidu A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL) International Journal of Computer Applications (0975 – 8887) Volume 68– No.16, April 2013

[13] Nirmala, M. and V. Palanisamy" AN EFFICIENT PREDICTION OF MISSING ITEMSET IN SHOPPING CART" Journal of Computer Science, 9 (1): 55-62, 2013 ISSN 1549-3636 2013

[14] U.Chandrasekhar, Sandeep Kumar. K2, Yakkala Uma Mahesh3, A Survey of latest Algorithms for Frequent Itemset Mining in Data Stream, International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-3 Number-1 Issue-9 March-2013.

[15],Sandeep Kumar. K,,Yakkala Uma Mahesh A Survey of latest Algorithms for Frequent Itemset Mining in Data Stream <a href="../se , INTERNATIONAL JOURNAL OF ADVANCED COMPUTER RESEAECH,2013,India

[16] Bolin Chen_, Min Li y, Jianxin Wang y and Fang-Xiang A logistic regression based 2/14/$31.00 © algorithm for identifying human disease genes Wu _ z978-1-4799-5669-2014 IEEE.

[17] M.A.NisharaBanuGomathy Disease Forecasting System Using Data Mining Methods 2014 International Conference on Intelligent Computing Applications 978-1-4799-3966-4/14 $31.00 © 2014 IEEE.

[18] Aida Mustapha et al. International Journal of Engineering and Technology (IJET) A Regression Approach for Forecasting Vendor Revenue in Telecommunication Industries Vol 6 No 6 Dec 2014-Jan 2015.

[19] GustiAyuPutriSaptawati 1, GustiNgurah Mega Nata Knowledge Discovery on Drilling Data to Predict Potential Gold Deposit, 2015 International Conference on Data and Software Engineering, 978-1-4673-8430-8/15/$31.00 ©2015 IEEE.

[20] Swati Gupta, A Multiple Regression Technique in Data Mining, International Journal of Computer Applications (0975 – 8887) Volume 126 – No.5, September 2015

[21] SheelaGole Frequent Itemset Mining for Big Data in social media using ClustBigFIM algorithm International Conference on Pervasive Computing (ICPC) -1-4799-6272-3/15/$31.00(c)2015 IEEE

[22] Swati Gupta, A Regression Modeling Technique on Data Mining, International Journal of Computer Applications (0975 – 8887) Volume 116 – No. 9, April 2015

[23] Ilayaraja M*, MeyyappanT , Efficient Data Mining Method to Predict the Risk of Heart Diseases through Frequent Itemsets4t h International Conference on Eco-friendly Computing and Communication Systems, ICECCS 2015 1877-0509 © 2015. Published byElsevier B.V.

[24] JaeKwonBaeandJinhwaKim,"A Personal Credit Rating Prediction Model Using DataMining in Smart Ubiquitous Environments, RESEARCH ARTICLE PUBLISHED IN FEBRUARY,2015

[25] Farhan Khan, Dariush Kari, IlyasAlperKaratepe, and Suleyman S. Kozat "UniversalNonlinear Regression on High Dimensional Data Using Adaptive Hierarchical Trees 2016 IEEE.

[26] ZhanfengWanga. . QingShib.eYoungjoLeec, Extended-T-process regression models Journal of Statistical Planning and InferenceVolume 189, (October 2017, Pages 38-60 © 2017 Elsevier)

[27]Sonja Greven, Fabian Scheipl"A general framework for functional regression modelling "irst Published February 28, (2017, sage journals, California)

MANAV RACHNA
Multidisciplinary Institution

Swami Vivekananda Library and Resource Centre
Manav Rachna University

SIMILARITY VERIFICATION CERTIFICATE

Title: A Novel approach for Enhancing privacy preservation in data Mining

Researcher: Pinki Singh

Registration Number: 2K16/PSE/08

Supervisor: Dr. Poonima Gupta

Department: Computer Science & Technology

This is to certify that the above mentioned Short Synopsis/Thesis was scanned for similarity index through Turnitin Software on 11 August with submission ID 987939245 and the similarity index was found to be 3 %. The complete report running into 23 Pages, 5317 Word count with File Size 4019 KB is being forwarded to the Research Supervisor/DRC/DRB for review.

Swami Vivekananda Library and Resource Centre, Manav Rachna University

11/8/2019

Date:

The complete report of the above Short Synopsis/Thesis has been reviewed and the similarity index is below the accepted norms as per the University Grants Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations 2018. The Short Synopsis/Thesis may be considered for further processing

Research Scholar

Pinki
11/8/2019

Research Supervisor

11/8/19

synopsis_11_august.docx

by

Submission date: 11-Aug-2019 07:35AM (UTC+0530)
Submission ID: 989079265
File name: synopsis_11_august.docx (100.53K)
Word count: 5397
Character count: 30814

Proposed Title of Thesis: A Novel Approach for improving Prediction in Data Mining.

Abstract: Data mining is the set of algorithms deliberate to find the hidden knowledge from the data sets, some of the popular techniques of data mining are prediction, sequential patterns, association, classification, clustering, and decision tree. Data mining techniques are categorized into two categories: Predictive and Descriptive techniques. Regression and classification are the predictive mining techniques. Association and clustering are the descriptive mining techniques. Classification and regression are used for forecasting. Regression algorithms are based on various regression model i.e. linear regressions, non-linear regression, multiple regressions, logistic regression, and probabilistic regression. Regression techniques have an important role in the prediction of continuous data, which can be frequent item stream data or time series data. It can be one dimensional or multidimensional stream data. Regression based techniques for prediction of stream data is frequent item prediction method (FIPM), frequent temporal pattern data stream (FTPDS), Sequence frequent stream data for plane regression stream (SFAPR).SFAPR overcome the problem of FIPM and FTPDS. Forecasting algorithms have important role in the area of traffic, weather forecasting, business analysis, share market, stock exchange, mobility, medical etc. In some area like drilling, stock exchange, weather forecasting prediction can help in increasing the revenues and performances. It can also be done by combining classification and regression techniques. In the proposed work forecasting of time series data sets with improved parameters is done. For preprocessing the data set, sliding window or classification algorithms are used. Then coefficients values for the regression model are identified to fit the regression model. It can be used to inspect the characteristics of the data.

Results: accurate prediction, reduction of standard errors, data cleaning. To improve the linear models quadratic, root or polynomial functions can be used. Evaluation of models are done to check the standard errors or residual plots so that it can help to avoid inadequate models and help to fitt the model for better results.

synopsis_11_august.docx

ORIGINALITY REPORT

Vivekananda Library and Resource Centre, Manav Rachna University
11/8/2019

Excluding matches less than 14 words

8%	4%	5%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Khan, Farhan, Dariush Kari, Alper Karatepe, and Suleyman Kozat. "Universal Nonlinear Regression on High Dimensional Data Using Adaptive Hierarchical Tree", IEEE Transactions on Big Data, 2016. Publication	1%
2	Submitted to Colorado Technical University Online Student Paper	1%
3	research.ijcaonline.org Internet Source	1%
4	Qing-Hua Li. "A Plane Regression-Based Sequence Forecast Algorithm for Stream Data", 2005 International Conference on Machine Learning and Cybernetics, 2005 Publication	1%
5	linknovate.com Internet Source	1%
6	ijarece.org Internet Source	<1%

7	halobi.com Internet Source	<1%
8	www.statisticshowto.com Internet Source	<1%
9	Karthikeyan, T., and K. Vembandasamy. "A Refined Continuous Ant Colony Optimization Based FP-Growth Association Rule Technique on Type 2 Diabetes", International Review on Computers and Software (IRECOS), 2014. Publication	<1%
10	Alberto Landi, Paolo Piaggi, Marco Laurino, Danilo Menicucci. "Artificial Neural Networks for nonlinear regression and classification", 2010 10th International Conference on Intelligent Systems Design and Applications, 2010 Publication	<1%
11	Submitted to Arab Open University Student Paper	<1%
12	Submitted to University of Melbourne Student Paper	<1%
13	"Mining Intelligence and Knowledge Exploration", Springer Nature America, Inc, 2013 Publication	<1%
	Submitted to Massey University	

14	Student Paper	<1%
15	Gusti Ayu Putri Saptawati, Gusti Ngurah Mega Nata. "Knowledge discovery on drilling data to predict potential gold deposit", 2015 International Conference on Data and Software Engineering (ICoDSE), 2015 Publication	<1%
16	wangqinhu.com Internet Source	<1%
17	Ilayaraja M., , and Meyyappan T.. "Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets", Procedia Computer Science, 2015. Publication	<1%
18	www.ijircce.com Internet Source	<1%
19	M., Suresh, Santhosh Kumar B., and S. Karthik. "A Load Balancing Model in Public Cloud Using ANFIS and GSO", 2014 International Conference on Intelligent Computing Applications, 2014. Publication	<1%
20	Kim, Hyunsung, Mike Sumption, Hojun Lim, and Edward Collings. "Evaluation of Mechanical Properties of Tubular Materials With Hydraulic	<1%

Bulge Test for Superconducting Radio Frequency (SRF) Cavities", IEEE Transactions on Applied Superconductivity, 2013.
Publication

21	www.theaccents.org Internet Source	<1%
22	Duck Jin Chai, Eun Hee Kim, Long Jin, Buhyun Hwang, Keun Ho Ryu. "Prediction of Frequent Items to One Dimensional Stream Data", 2007 International Conference on Computational Science and its Applications (ICCSA 2007), 2007 Publication	<1%
23	Jae Kwon Bae, Jinhwa Kim. "A Personal Credit Rating Prediction Model Using Data Mining in Smart Ubiquitous Environments", International Journal of Distributed Sensor Networks, 2015 Publication	<1%
24	Thomas Kurner. "Application of Classification and Regression Trees for Paging Traffic Prediction in LAC Planning", 2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring, 04/2007 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		